# Parallel Information Retrieval on A PC-Cluster Using Vector Space Model

Mr. Tawa Khampachua

**Tawa.K@Student.chula.ac.th**

**Software Engineering Laboratory (SE)**
**Department of Computer Engineering**
**Chulalongkorn University**

# Agenda

**Problem and Motivation**

**Proposed Solution**

**Fundamental Theories**

**Design and Implementation**

**Experimental Results**

**Conclusions**

# Problem Statement and Motivation

## Problem Statement

- Large-scale text databases.
- Limitation of computing resources.

## Motivation

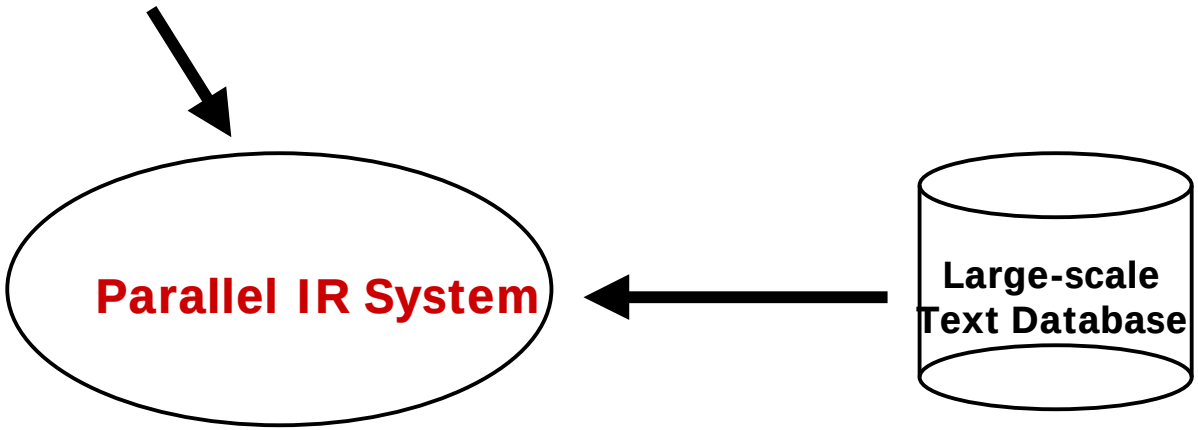
- Emerging of parallel and distributed computing on Beowulf PC-Cluster.

# Proposed Solution

**Parallel computing**
**(On cluster of PCs)**

**+**

**Information Retrieval (using VSM model)**



# Information Retrieval

## What is Information Retrieval

**Information Retrieval** is the process of identifying and retrieving relevant

documents based on a query consists of three basic steps;

- a document representation.
- a query representation.
- a similarity measurement between both of documents and queries.

# Information Retrieval

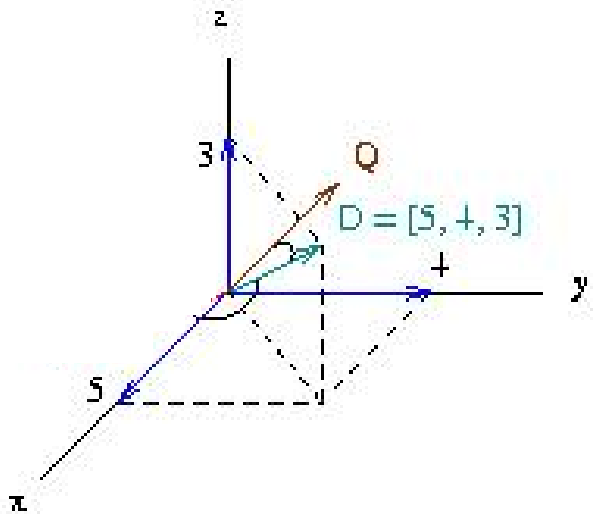
**The common classic models in information retrieval**

- **Exact match** : based on text pattern and Boolean search techniques.
- **Inexact match** : based on statistic search techniques.

# Information Retrieval

## Vector Space IR Model

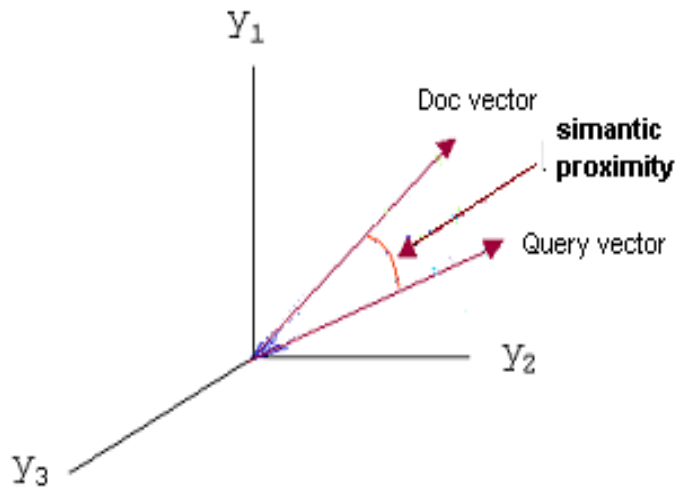
- Representing a document with a vector.
- Keeping document semantics with the length and direction of the vector.



# Information Retrieval

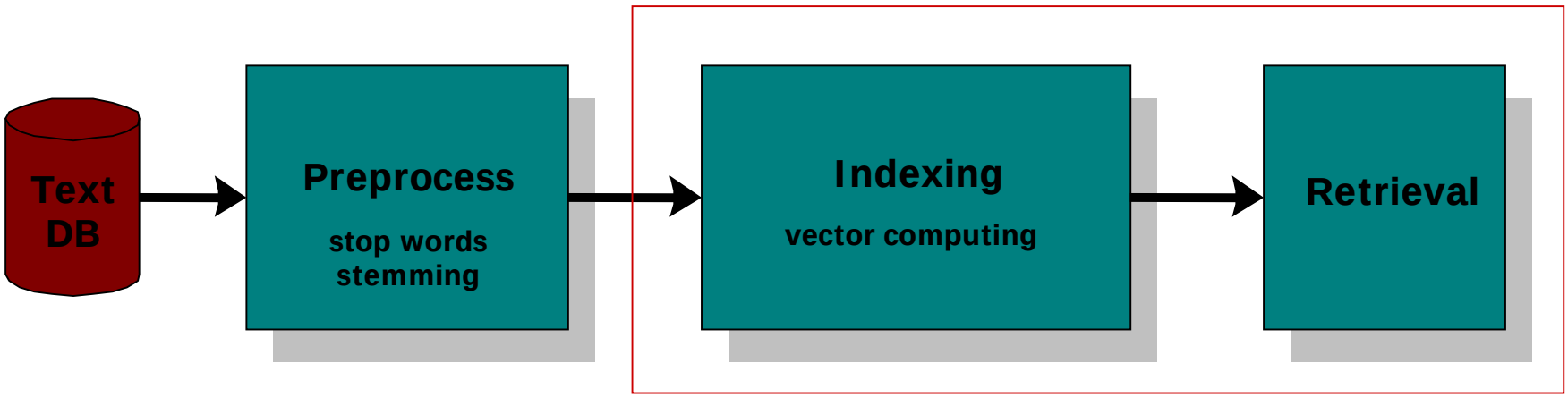
## Retrieval Phase

- Query vector formulation.
- Similarity measures ; determining similarity of documents with simple vector operations (dot product).



# Information Retrieval

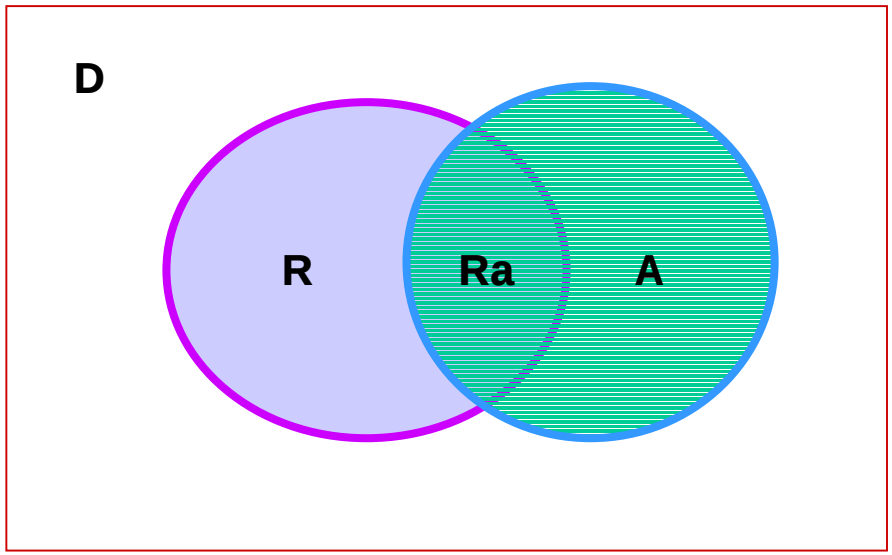
## IR system functional diagram




- Stop words : eliminate non useful words, e.g., a , an, the , etc.
- Stemming : reduce vocabulary size.
- Doc derive : construct document vectors.
- Retrieval : similarity computation.

# Information Retrieval

## Retrieval Evaluation




- Recall = Ra / A
- Precision = Ra / R

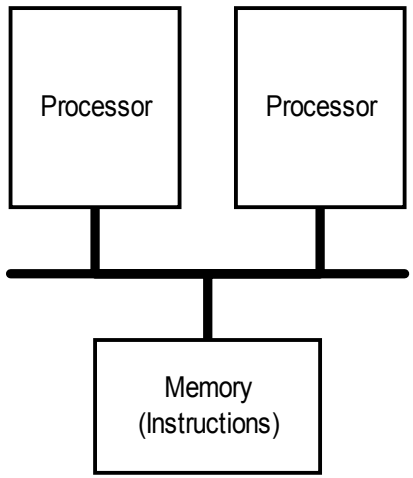
# Parallel and Distributed Computing

## Parallel computing concept

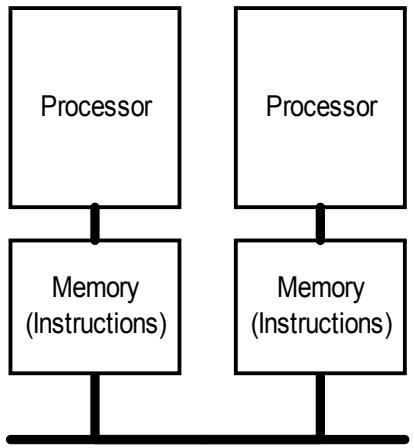
# Decomposing a large problem into more parts.
# Processing each part in parallel by more processors.
# Reduce the time required to solve the problem.
# Scale up the larger problems.

# Parallel and Distributed Computing

## SIMD and MIMD




SIMD Architecture




MIMD Architecture

# Parallel and Distributed Computing

## Performance Evaluation

- **Speedup** $S_p = \frac{T_1}{T_p}$
- **Efficiency** $E_p = \frac{S_p}{p}$

# Problems and Solution

**Problems**

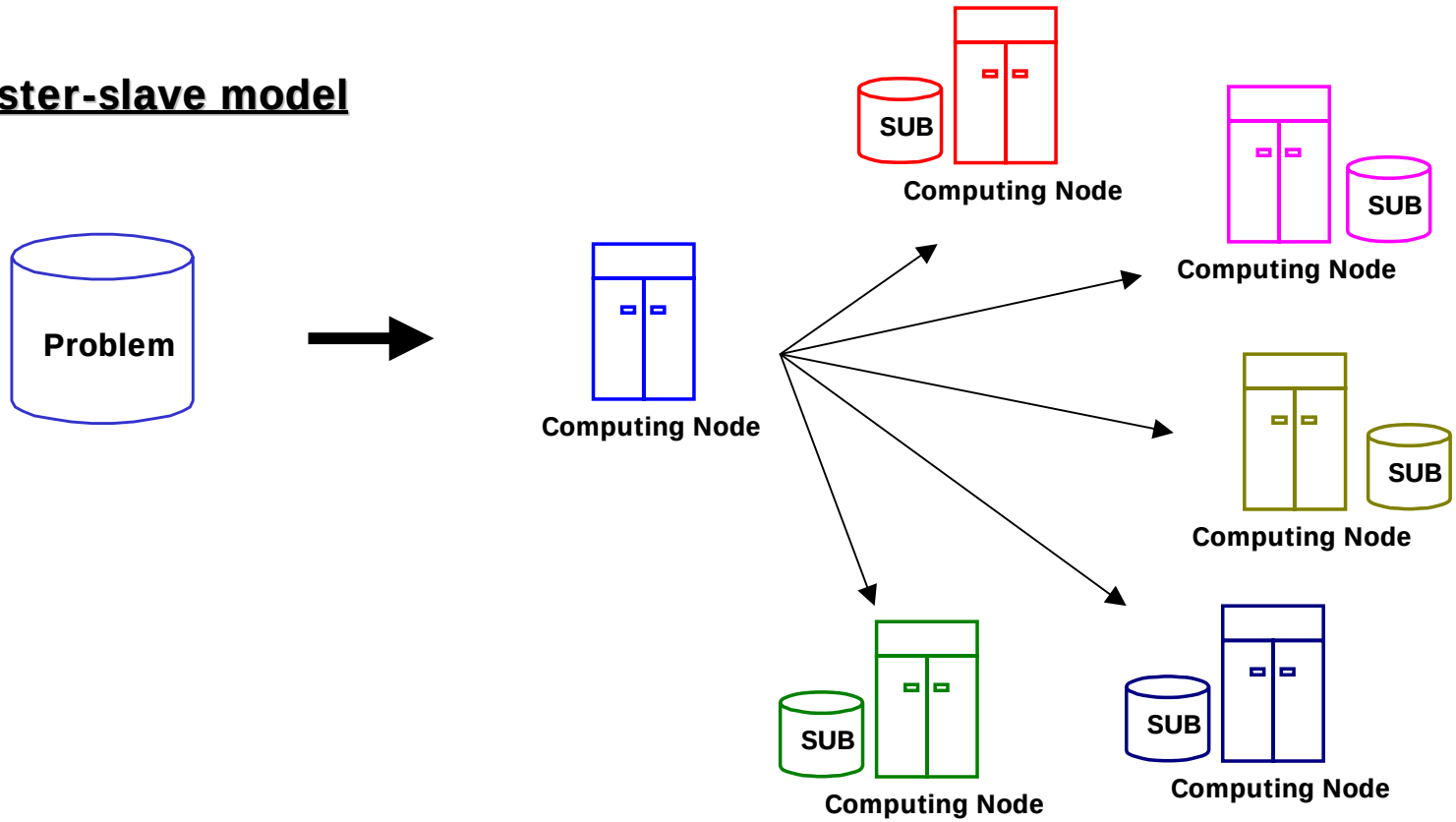
- A large-scale text database.
- A sequential system have a resources limitation.

**Solution**

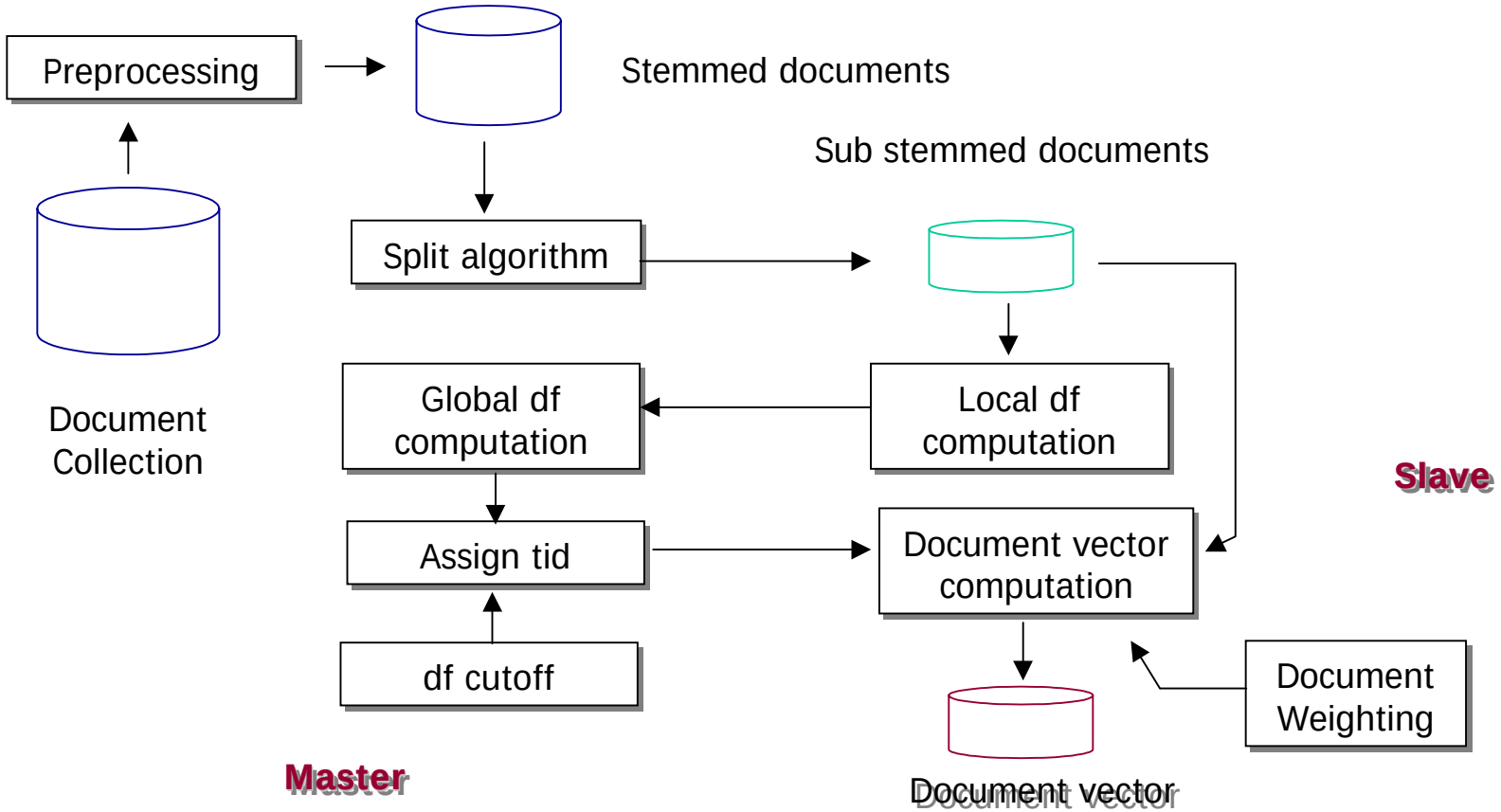
- Apply parallel computing with IR system.
- Split a large text database into small pieces and conquer with computing nodes .

# Parallel indexing using master-slave model

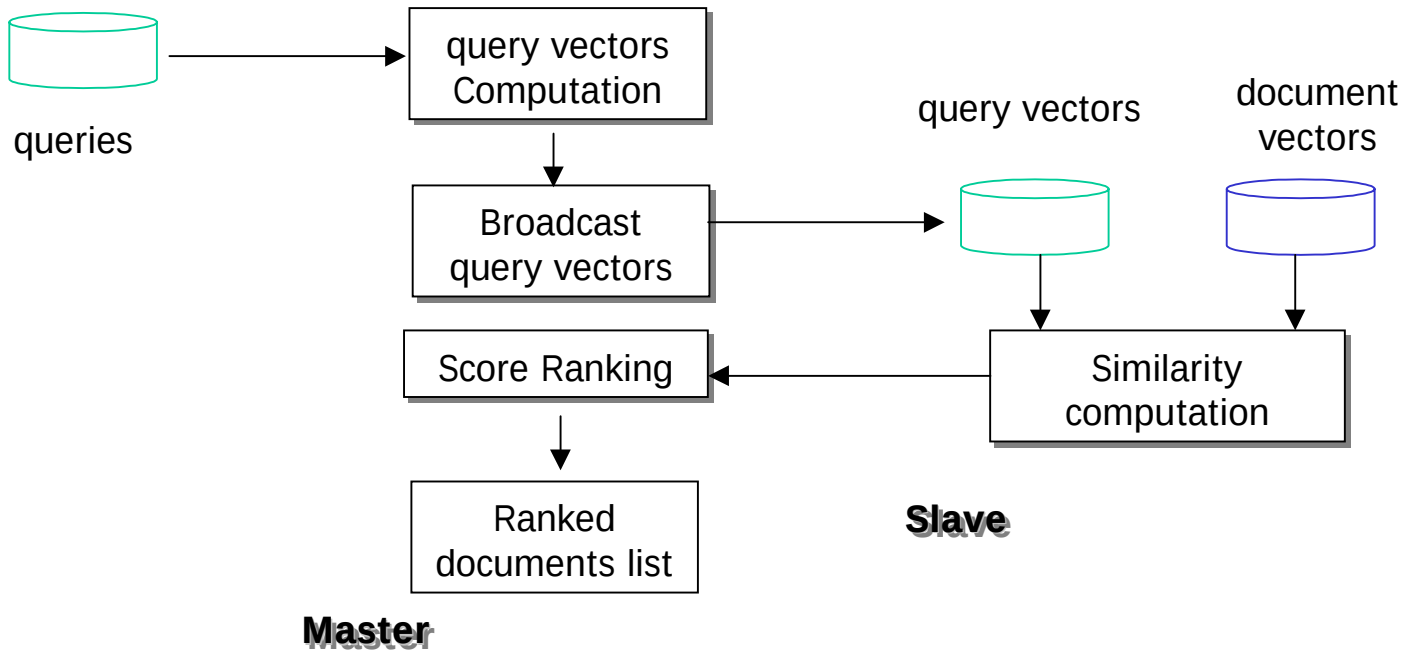
## Master-slave model



# Parallel indexing functional diagram



# Parallel retrieval functional diagram



# Materials and Configuration

## Collections

- Test collection, Cranfield and TREC-8 (2GB).

## Testing

- Retrieval performance (Precision).
- Speed-up performance.

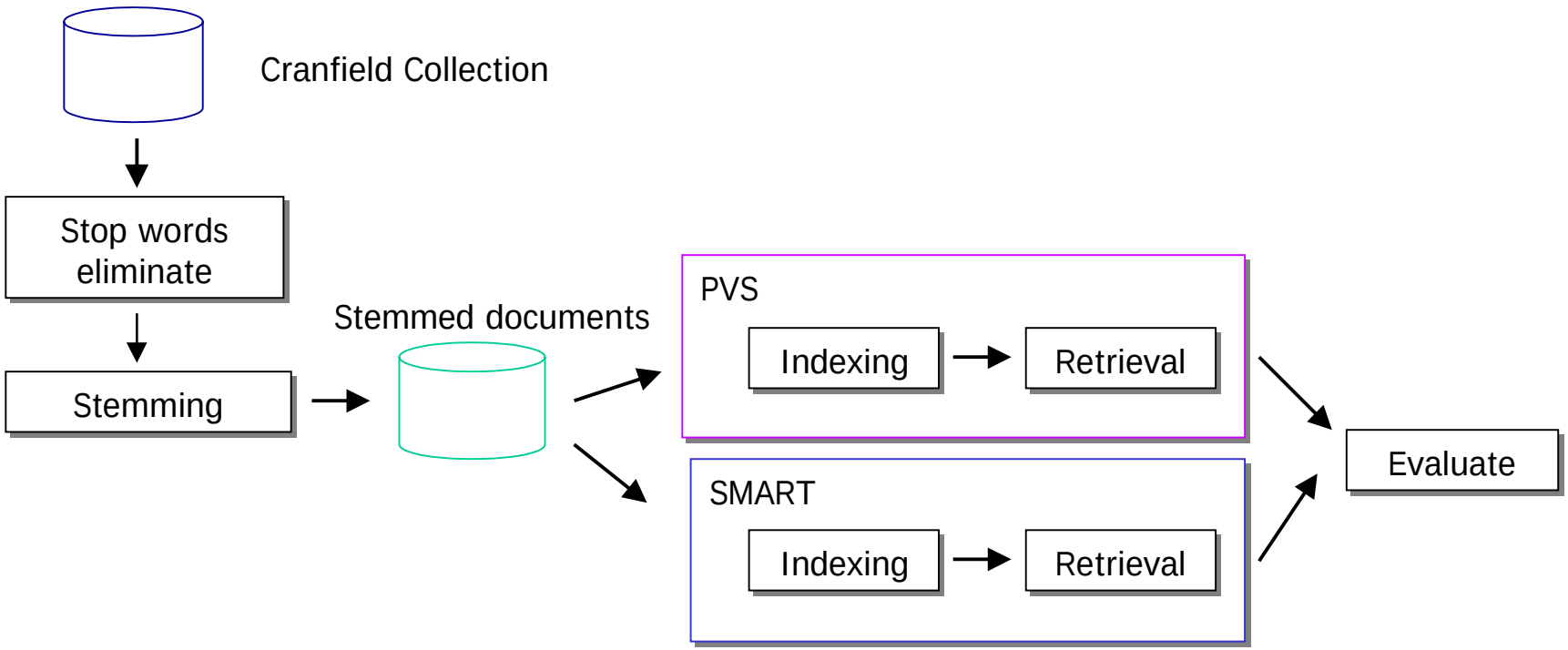
# Retrieval performance

## Cranfiled Collections

| | |
|---|---|
| Size in Mbytes | 1.6 |
| Number of documents | 1400 |
| Number of queries | 225 |
| Maximum terms/document | 225 |
| Minimum terms/document | 18 |
| Average terms/document | 88 |
| Total terms found | 125822 |
| Total unique terms | 3959 |

# Retrieval performance

## Test Diagram



# Retrieval performance

## Retrieved Result

| Cranfiled Collection | PVS | SMART |
|---|---|---|
| Retrieved | 22500 | 22500 |
| Relevant | 1838 | 1838 |
| Rel_retrieved | 1295 | 1295 |
| Average precision | 0.3531 | 0.3531 |

# Speed-up performance

## TREC Collections

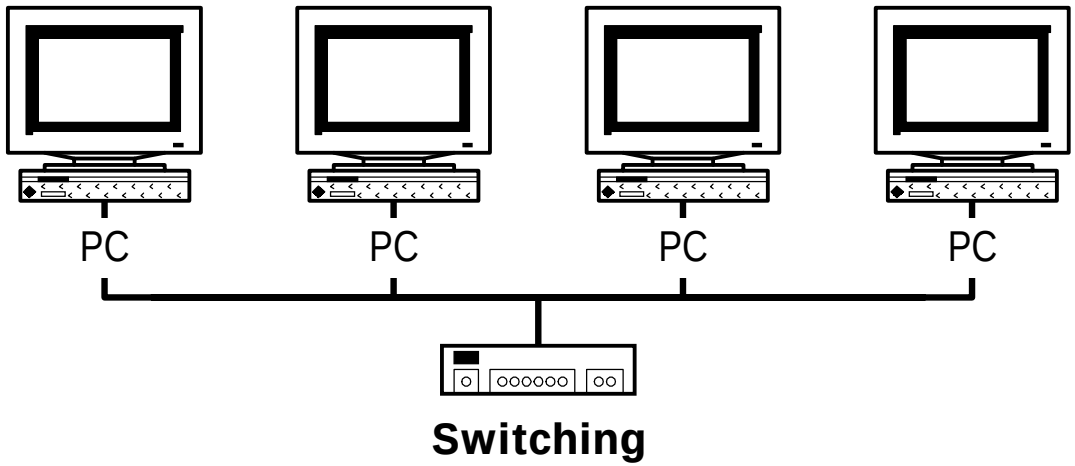
| Collections | LATIME | FR | FT | FBIS |
|---|---|---|---|---|
| Total size (MB) | 498 | 414 | 591 | 493 |
| Total documents | 131896 | 55630 | 210158 | 130471 |
| Total words | 70494552 | 38154925 | 84433831 | 67804697 |
| Unique words | 285643 | 157939 | 385795 | 321394 |

# Speed-up performance

## Test Collections

| Computing nodes | Size of data (MB) | Documents |
|---|---|---|
| 1 | 1,969 | 1,689,272 |
| 2 | 985 | 840,636 |
| 4 | 492 | 420,318 |
| 8 | 246 | 210,158 |
| 16 | 123 | 105,079 |

# Speed-up performance



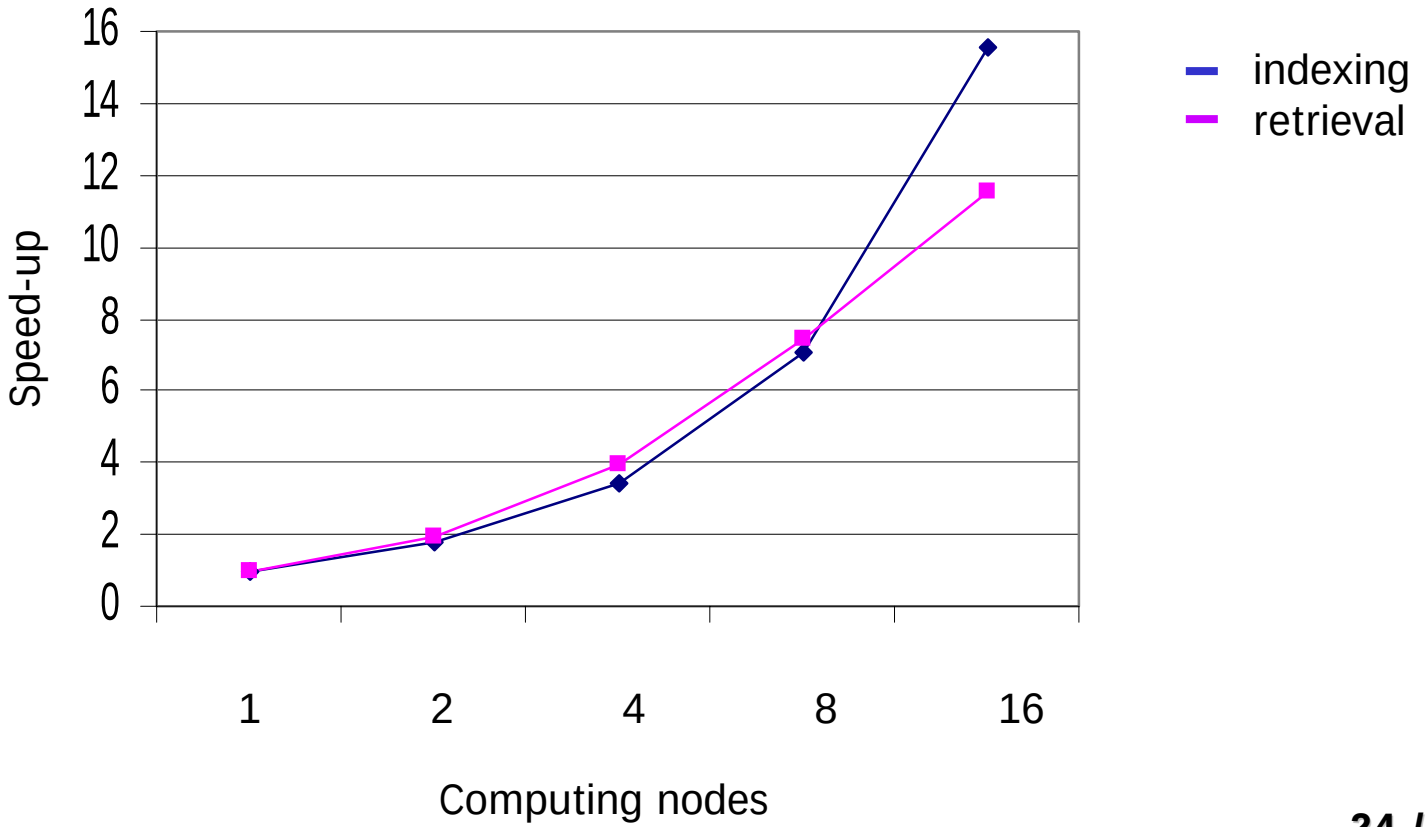

## Beowulf PCs cluster

16 machines of Athlon 700MHz with 128MB RAM and IDE 10 GB.

Each of node is connected with 100Mbps Switching.

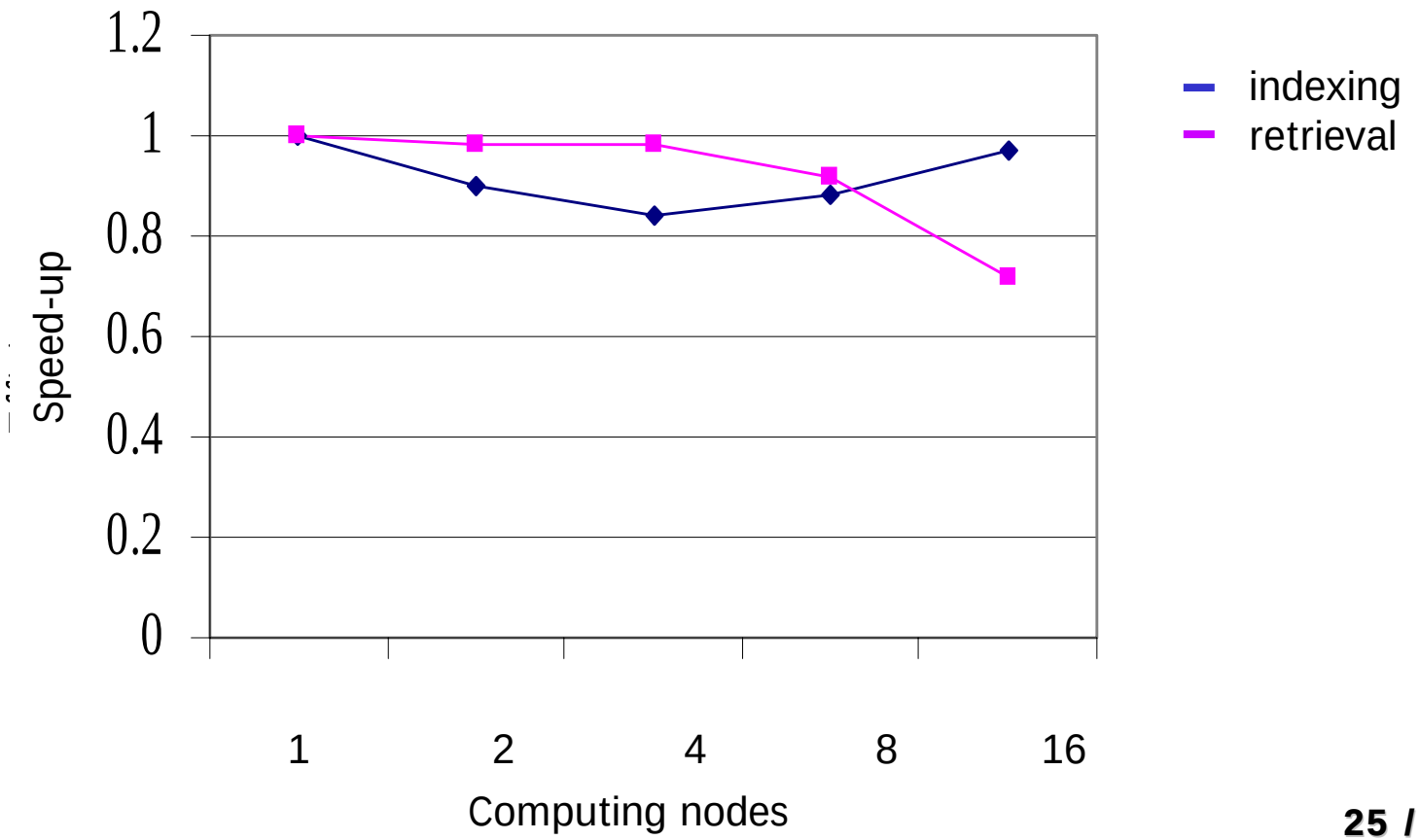
## Experimental Results - 8

# Speedup curve

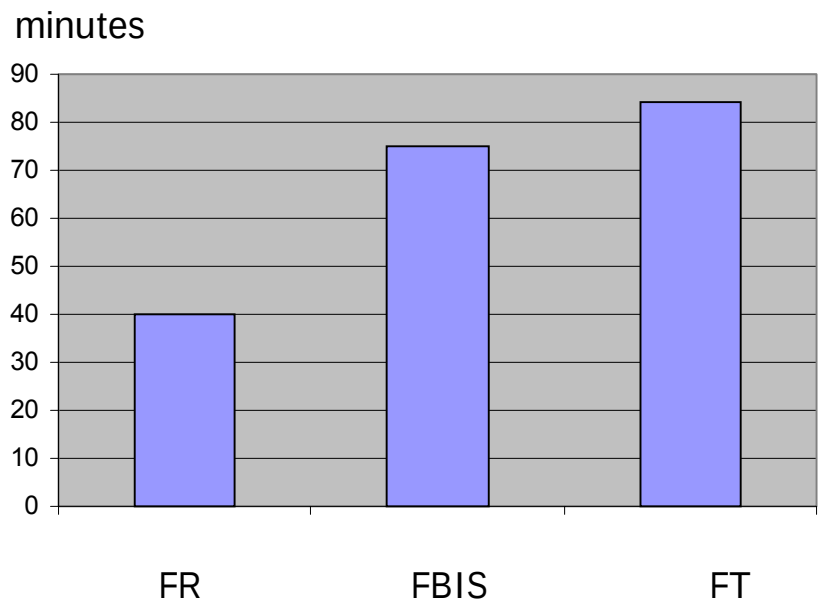


## Experimental Results - 9

# Efficiency of indexing

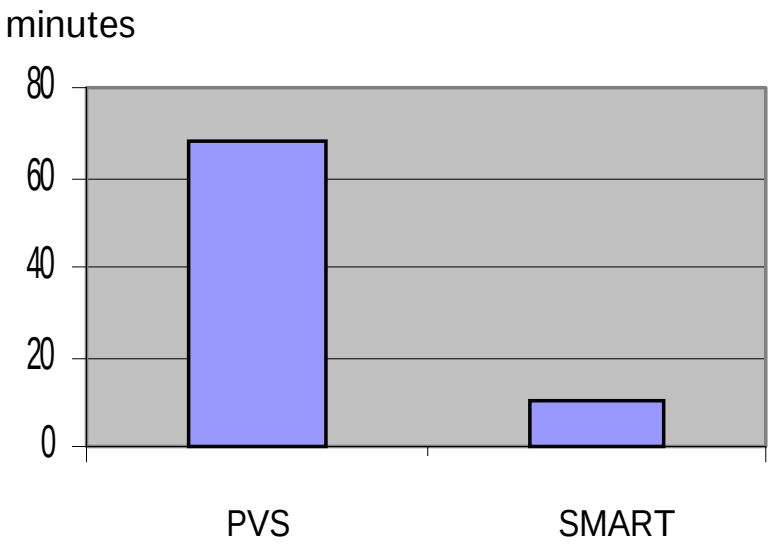


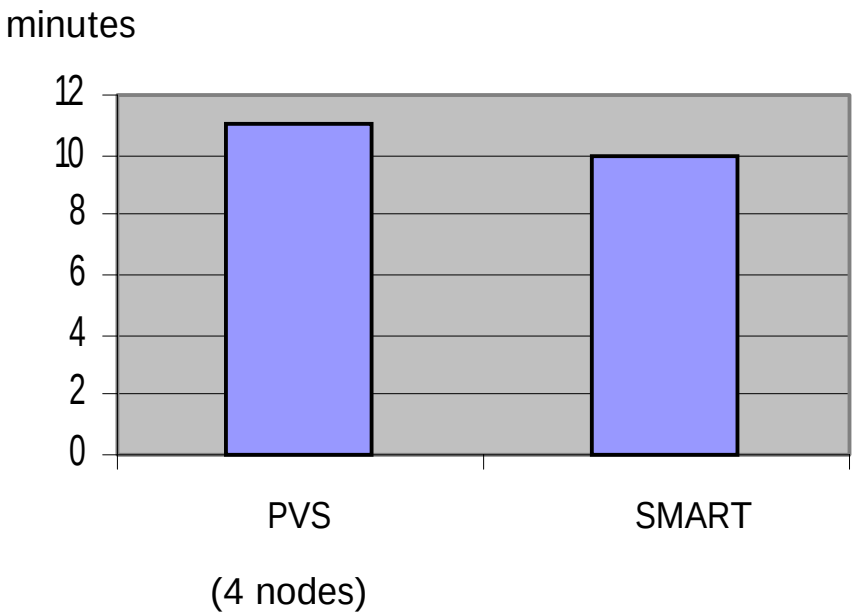
# Effect of Data Collection




Computing time

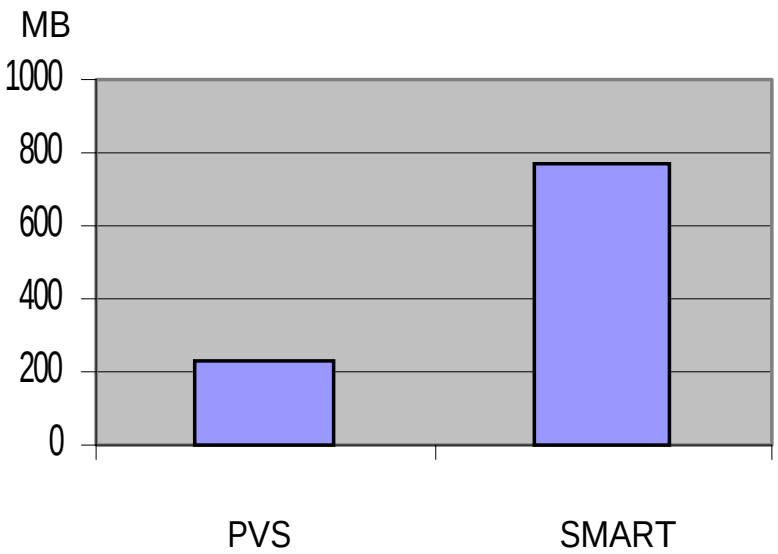
# Comparison to SMART



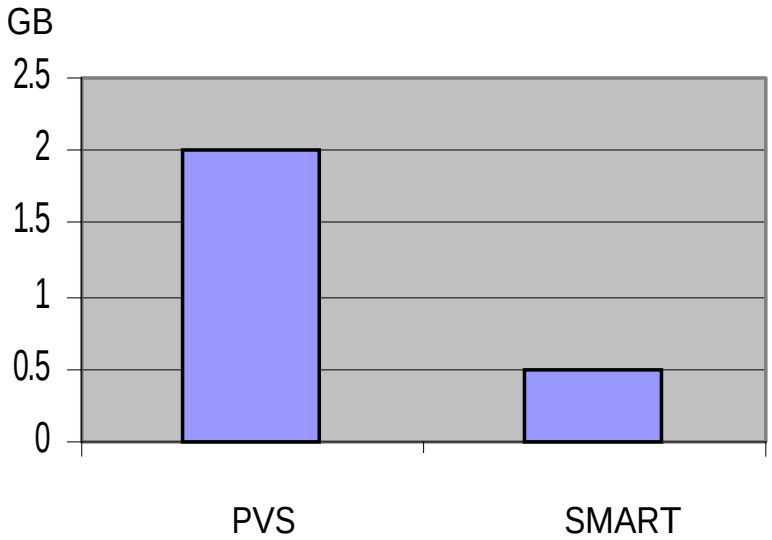




Computing time

# Comparison to SMART




Space




Maximum Data size

# Conclusion

**Problem**

- Large-scale text databases.
- Limitation of computing resources.

**Propose**

- Apply Parallel computing to information retrieval system.
- Using low-cost PCs cluster for implementation.

**Consequence**

- Parallel computing technique can help ;
  - to reduce an execution time.
  - to solve resources limitation.

# Conclusion

## Perspective

- Apply Inverted file for improve speed of retrieval process.
- Optimize Parallel algorithms to reduce computing time.
- Add new functions and retrieval techniques to improve precision of system.